A workshop for MATH1530 tutors.

*The goals of the workshop are:*

1. *To inform MATH1530 tutors on the differences in the content of the two special sections of MATH 1530 (Section 88, the honors section & Section 10, the pre-health section) with respect to the standard content of the other sections.*
2. *To review the extra material covered in these two sections*
3. *To providing the tutors with a basic set of instructions in R to perform the usual calculations and obtain statistical graphs with R that are studied in any regular section of MATH1530 and the special sections as well.*

I **About sections 10 and 88**

1. **Who are in sections 10 and 88 ?** How many students are there (36+8=44 this semester) Since when we do this? About 4 years , but the composition of the section has been different, this is the first time we have the two target groups together, so many times I have had just regular students there.
2. What do we do differently? The content originated in the material we prepared for the Symbiosis project (HHMI funded, 2006-2010) but later evolved in a stand-alone stats course. We introduce inference right from the beginning of the course, for that we need to include some randomization tests & bootstrapping. We also give importance to the binomial distribution and its use in testing hypothesis about a proportion, that allows us to introduce the concept of power in an easy way. Conditional probability in the context of medical diagnosis is seen including all the vocabulary of ‘sensitivity’, ‘false positive’, PPV etc, we use probability trees to apply the Bayes rule. We still do the classical inference with the t etc. toward the end of the semester.
3. List of topics

**List of topics to cover:**

1. Definitions: Hypotheses, variables, parameters and statistics (Ch. 1- sections 1.1-1.4).
2. Producing data using surveys and experiments; preparing data files (Ch. 1- sections 1.5-1.8).
3. Graphs and statistics for one quantitative variable (Ch. 2, Sec. 2.1-2.11).
4. Graphs and statistics for two or more quantitative variables: Scatter plots and correlation (Ch. 2, Sec. 2.12-2.13).
5. Graphs and statistics for categorical variables (Ch. 2, Sec. 2.14-2.15).
6. Randomization test (Ch. 3, Sec. 3.1).
7. Bootstrapping to build confidence intervals (Ch. 3, Sec. 3.2).
8. Introduction to probability and the binomial distribution (Ch. 4).
9. Testing hypotheses with the binomial distribution (Ch. 5).
10. Conditional probability and Bayes Rule using probability trees (Ch. 6, Sec. 6.1-6.4.1).
11. Normal and Chi-square distributions (Ch. 9, Sec. 9.1,9.3, 9.5).
12. Checking models and assumptions: Chi-square tests and test for normality (Ch. 10).
13. Sampling distributions and confidence intervals for the mean (Ch. 11, Sec. 11.1-11.2).
14. Testing hypotheses for means (t-tests) (Ch. 11, Sec. 11.3, 11.5).
15. Large sample inference for proportions (Ch. 11, Sec 11.4.1-11.4.4).
16. Introduction to regression (Ch. 12, sections 12.1-12.6).

We try to keep the evaluation system as close as possible as the rest of the sections of MATH1530

Exam 1 (Topics 1-10A) 175 Thursday, October 9

Exam 2 (Topics 10A-16) 175 Tuesday, November 25

Quizzes (20 points each) 200 (see attached calendar)

Labs(4) \* 100 (see attached calendar)

Final exam (comprehensive: 200 Tuesday Dec 9 at 8am

Final data analysis project 100 Tuesday December 2

Online practice quizzes\*\* 50 (each one has its own deadline)

TOTAL\*\*\* 1000

+ bonus points for attendance (50 max)

1. Calendar **TENTATIVE CALENDAR – MATH 1530 – SECTIONS 010– FALL 2014 (once in a while we make a small adjustment)**

|  |  |  |
| --- | --- | --- |
| WEEK | TUESDAY (classroom: Gilbreath 304) | THURSDAY (computer Lab: Gilbreath 305/306) |
| 1 | 8/26 Read Chapter 1Topic 1 : Chap1-definitions,hypothesesDiscuss exercises2 &4 in Chapter 1 | 8/28 Read Chapter 1Topic 2 : Chap1-surveys & experimentsAn introduction to R |
| 2 | 9/2 Read Sections 2.1-2.11**Finish Topic 2**Topic 3: Graphs and summaries for one quantitative variable**Quiz # 1 on Topic 1 (hypotheses , definitions etc)** | 9/4 Read Sections 2.1-2.11Topic 3 (continuation) Learning how to produce plots and to calculate statistics with the computerAn introduction to Minitab **Assign Lab 1 on Chapter 2 (descriptive statistics)** |
| 3 | 9/9 Read Sections 2.12-2.13Topic 4: Scatterplots, correlation**Quiz # 2 on topic 2 (Surveys and experiments)** | 9/11 Read Sections 2.14-2.15Topic 4 (scatter plots and correlation with statistical software)Topic 5: Graphs and summaries for categorical variables  |
| 4 | 9/16 Read Section 3.1Topic 6 Randomization test **Quiz #3 on topic 3 (descriptive stats)** | 9/18 Read Section 3.2Topic 7 The Bootstrap methodPracticing Randomization tests and bootstrap with the computer.**Quiz # 4 on topics 4 & 5 (correlation, two-way tables etc)** |
| 5 | 9/23 Read Chapter 4Topic 8 : Probability and Binomial distribution**Quiz # 5 on topics 6 and 7 (randomization test & bootstrap)** | 9/25 Read Chapter 4Topic 8: Probability and Binomial distribution**Lab 1 on Chapter 2 is due** |
| 6 | 9/30 Read Chapter 5Topic 9 Testing hypothesis with binomial | 10/2 Read Chapter 5Topic 9 Testing hypothesis with binomial & concept of power. **ASSIGN FINAL PROJECT Quiz # 6 on topic 8 (Probability &binomial)** |
| 7 | 10/7 Read Chapter 6**Quiz # 7 on Topic 9 (Test for p using the binomial)**Topic 10 A -Conditional probability from 2-way tables and formula P(B/A)=P(A∩B)/P(A)  | 10/9 **EXAM 1 on Topics 1-10A (from the beginning up to 2-way tables)** |
| 8 | 10/14 FALL BREAK | 10/16 Topic 10B Conditional Probability: Bayes rule using probability trees**. Assign Lab 2 on topic 10 (Conditional probability)** |
| 9 | 10/21 Read Sections 9.1-9.3Topic 11: Normal distribution**Quiz # 8 on Topic 10 (A & B) (conditional probability)** | 10/23 Read sections 9.3 & 9.5Topic 11: Normal and Chi-square distributions**Lab 2 on conditional probability is due** |
| 10 | 10/28 Read Chapter 10Topic 12: Chisquare tests (goodness of fit, independence , homogeneity)**Quiz #9 on Topic 11 (Normal & Chi-square dist.)** | 10/30 Read Chapter 10Topic 12: finish Chisquare tests & normality tests  **Assign Lab 3 on topic 12 (Chisquare and normality tests)** |
| 11 | 11/4 Read Sections 11.1-11.2 Topic 13: Sampling distributions, confidence interval for the mean, sample size | 11/6 Read Sections 11.1-11.2Finish Topic 13 **Lab 3 is due**Start Topic 14: Testing hypothesis for means |
| 12 | 11/11 Read sections 11.3-11.5Topic 14: testing hypothesis for means**Quiz #10 on Topic 13 (sampling dist. & confidence int.)** | 11/13 Read sections 11.3 &11.5Topic 14: Testing hypothesis for means**Assign Lab 4 on topic 14 (Inference for means)** |
| 13 | 11/18 Read section 11.4 Topic 15A. Large sample inference for proportions (confidence interval & sample size ) **Quiz # 11 (** on topic 14 to prepare for the test) | 11/20 Read section 11.4 Topic 15B. Large sample inference for proportions (test of hypotheses)**Lab 4 is due**  |
| 14 | 11/25 Read sections 11.4.1-11.4.4 Finish Topic 15 **EXAM 2 (on Topics 10A-15A)** | 11/27THANKSGIVING |
| 15 | 12/2 Read sections 12.1-12.6 Topic 16: Introduction to Regression **FINAL PROJECT IS DUE** | 12/4 Read sections 12.1-12.6Finish Topic 16 Introduction to Regression  |

II **A short review of some of the extra material**

1. **Hypotheses & data**

Hypotheses are important in the scientific method, scientific hypotheses cannot be ‘proved’ to be truth, they are only falsifiable (so we don’t ‘accept an hypothesis’ we only ‘reject’ or ‘not reject’ it. We make decisions about hypotheses based on data. We emphasize the definitions of individual, population, variable (quantitative or categorical), parameter and statistic; give the students a story and ask to identify those in the story. Examples

A public health researcher wants to know what proportion of the population of Tennessee, 18 years or older, can be considered current smokers. He designs a survey that has several questions and using the answers to those questions he will determine for each individual he/she can be considered a current smoker or not. He selects a random sample of 2000 residents of Tennessee, 18 years or older. Out of the 2000, 480 are classified as current smokers

A graduate student is conducting a research on the performance of freshmen at a given college. He organizes a survey for which he selects a random sample of 200 freshmen. He wants to know what is the mean number of hours that freshmen spend doing homework during the third week of the semester. One of the questions he asks on Monday of the 4th week is ‘How many hours did you spend doing homework last week?’ He collects the answers of the 200 students and the mean number of hours for those 200 students is 6.5 hours.

Statistical hypothesis are written in terms of parameters, we ask them to write the statistical hypotheses for a given story. Examples:

A biologist hypothesizes that adult male flesh flies are more prone to exhibit aggressive behavior than females of the same species and age. He plans to observe, during 2 hours, a group of 30 male flies and 30 female flies, all of them 4 days old. He will record for each one of the flies whether they exhibit aggressive behavior or not during that period of time. **Write the null and alternative statistical hypotheses**

A biologist hypothesizes that adult male flesh flies are more aggressive than females of the same species and age. He plans to observe, during 2 hours, a group of 30 male flies and 30 female flies, all of them 4 days old, and record for each one of the flies the total amount of minutes that they spend engaged in aggressive behaviors . **Write the null and alternative statistical hypotheses**

We give them an intuitive idea of the ‘p-value’ as the probability of getting the results we got in a survey or experiment when the null hypothesis is true, and that we reject Ho if pvalue is small (because if it is very unlikely that we get what we got when Ho is true, and we got it , well, maybe Ho is not true). We also tell them that there are several ways to calculate and approximate the p-value depending on the circumstances and that we will learn them during the semester.



1. Randomization test

We introduce the randomization test to compare two populations using a hands on activity and then we apply it by using R-code that mimics the hands-on activity. To apply it to other data sets they simply have to change the data. Randomization tests are used in introductory statistics courses in several parts of the USA and there is even a user friendly software statkey developed by the authors of ‘Unlocking the power of Statistics’ (by 5 members of the Lock family) that they generously put it in the internet to be used by anybody.

The phytate story



 

The data



 

The activity







Applying the randomization test

R-code available at http://faculty.etsu.edu/seier/RcommCh3.txt



1. Bootstrapping

The rhododendron story



The sample



Sampling with replacement (from the sample, using the same n) and ‘bootstrap replicates’ of the statistic

 

The idea of the percentile confidence interval

 

In this table it is very easy for the students to understand why when you increase confidence the interval becomes wider.

There are packages in R to do boostrapping, you can also use statkey, but we use our code on

<http://faculty.etsu.edu/seier/RcommCh3.txt>

The beauty of bootstrap, you can build confidence intervals for things other than the mean without need of formulas, just change the statistic in the code



Using confidence intervals to test hypotheses



1. **Testing hypotheses using the binomial distribution including the notion of power**

Most intro-stats books have a warning against using the normal approximation to test Ho: p=po when the sample size is small but not always provide an alternative method that happens to be the so called ‘exact’ test for proportions based on the binomial distribution. However, Minitab has it as an option.

We start with a hands on activity to test Ho: p=0.5 Ha: p>0.5 where p is the proportion of red in a two-color plastic chip. We ask the students to toss a chip 10 times and count in how many of them the red color comes up. (we could do the same with a quarter). Then we ask them to calculate the probability of that number of a more extreme one happening by chance when p=0.5 . That is the p-value of the test

We also do other examples such as the story of the ducks

For one sided alternatives, assuming there were 9 successes in 10 trials



For two-sided alternatives assuming there were 8 successes in n trials



We do other examples with other values of n or p. For example

assuming there were only 3 successes



More about α.
We tell them that fixing a value of α is equivalent to defining a decision rule. For example if I am testing Ho:p=0.5 vs Ha: p>0.5 with a sample n=10, and I say that I will reject Ho if I find 8 or more successes, what the value of α? From there we go to define power , and generally one of them asks where is the other type of error.

The binomial table is a good instrument for a discussion of these concepts and bringing them from a definition to something more tangible. Instead of telling them rules of how to calculate them we discuss the concepts and push them into thinking how to calculate them.



We briefly discuss power in relation to sample size.



1. Conditional probability in the context of medical tests both from two-way tables and probability trees. We introduce the vocabulary of sensitivity, specificity, -, false +, PPV and NPV

We start as usual calculating conditional probabilities from 2-way tables. With respect to diseases we get familiar with the 6 conditional probabilities













We can calculate all those conditional probabilities provided the two-way table reflects the prevalence of the disease or condition in a population (or a subpopulation of interest such as a population of high risk etc)

Some examples are:

For the USA population around 1998



For high risk pregnancies



From a table like this that reflects only the results of an experiment about the test but not the prevalence of the disease in the population , we could calculate the sensitivity, specificity, prob of false + and prob of false -, but not PPV or NPV



However in the medical / public health literature the information usually does not come in the form of a table but of sensitivity, specificity and prevalence, then probability trees help to apply the Bayes rule to calculate reverse probabilities such as PPV and NPV

For the high risk population of pregnancies, calculating the probability of Down Syndrome given that the result of the test was negative (Normal Marker)





We discuss how to calculate PPV=P(melanoma/test + ) and NPV=P(no melanoma/test -) depending on the prevalence of melanoma in the subpopulation we are talking about.